(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 March 2002 (21.03.2002)

PCT

(10) International Publication Number
WO 02/23390 A2

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/EP01/10399

(22) International Filing Date: 9 August 2001 (09.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
00480085.0 12 September 2000 (12.09.2000) EP

(71) Applicant *(for all designated States except MC)*: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, NY 10504 (US).

(71) Applicant *(for MC only)*: **COMPAGNIE IBM FRANCE** [FR/FR]; Tour Descartes, 2, avenue Gambetta, F-92066 Paris La Défense Cedex (FR).

(72) Inventors: **LAMBERTON, Marc**; 981, route de St Jean, F-06600 Antibes (FR). **LEVY-ABEGNOLI, Eric**; 67, Ancien Chemin de la Lanterne, F-06200 Nice (FR). **THUBERT, Pascal**; Les Jardins d'Elise, 60, avenue des Poilus, F-06140 Vence (FR).

(74) Agent: **DE PENA, Alain**; Compagnie IBM France, Direction de la Propriété Intellectuelle, F-06610 La Gaude (FR).

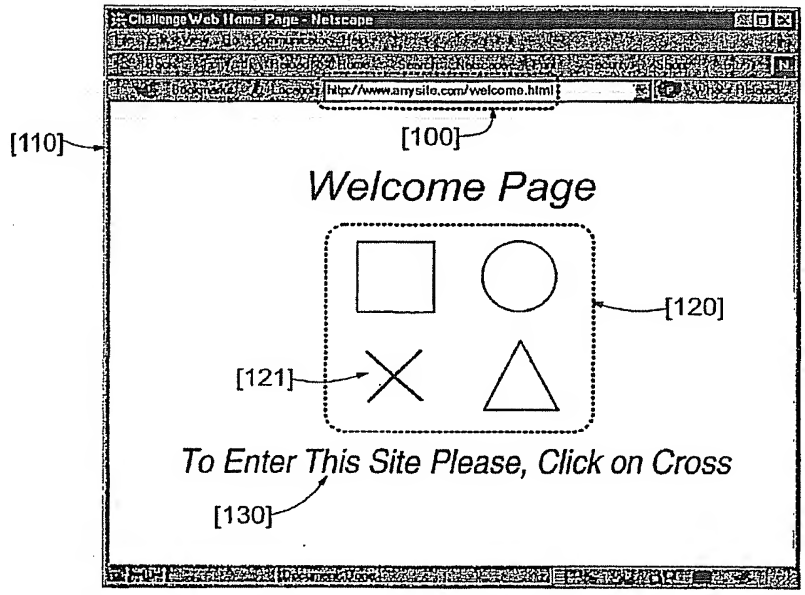
(81) Designated States *(national)*: CA, CN, JP, KR, SG.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

**Published:**
— *without international search report and to be republished upon receipt of that report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: SYSTEM AND METHOD FOR IMPLEMENTING ROBOT PROOF WEB SITE




(57) Abstract: The invention allows to prevent robots from browsing a Web site beyond a welcome page. When an initial request from an undefined originator is received Web site responds to it with a welcome page including a challenge. Then, on receiving a further request from the undefined originator Web site can check whether the challenge is fulfilled or not. If fulfilled the undefined originator is assumed to be a human being and authorized to go on. If challenge is not however fulfilled the undefined originator is assumed to be a robot in which case site access is further denied. The invention prevent Web site contents from being investigated by robots while not requiring users to have to log on.

WO 02/23390 A2

# SYSTEM AND METHOD FOR IMPLEMENTING ROBOT PROOF WEB SITE

## Field of the Invention

The present invention relates to the Internet and more particularly applies to those of the World Wide Web (WWW) sites that, while welcoming human beings, want to exclude robots from visiting and gathering information from them.

## Background of the Invention

WWW robots, also called Web Wanderers, Web Crawlers or Web Spiders and often just referred to as bots (bot is short for robot), are programs devised to automatically traverse the hypertext structure of the Web thus, having retrieved a document, can recursively retrieved all the linked pages. Especially, this is the case of the numerous search engines and their robots which roam the World Wide Web finding and indexing content to add to their databases. Although most robots provide a valuable service this has developed a certain amount of concern amongst Web site administrators about exactly how much of their precious server time and bandwidth is being used to service requests from these engines. If the majority of robots are well designed, are professionally operated and cause no problems, there are occasions where robots visiting Web servers are not welcome. Sometimes because of the way robots behave. Some may swamp servers with rapid-fire requests, or retrieve the same files repeatedly. If done intentionally this is a form of Denial of Service (DoS) attack

although this is more often just the result of a poor or defective robot design. In other situations robots traverse parts of WWW servers that are not suitable for being searched e.g., contain duplicated or temporary information, include large documents or e.g., CGI scripts (CGI is a standard for running external programs from a World-Wide Web HTTP server). In this latter case and in similar situations, when accessed and executed, scripts tend to consume significant server resources in generating dynamic pages thus, slow down the system. In recognition of these problems many Web robots offer facilities for Web site administrators and content providers to limit what the robot is allowed to do. Two mechanisms are provided. One is referred to as the 'Robots Exclusion Protocol' even though it is not really an enforced protocol but was a working document discussed as an Internet-Draft by the Internet Engineering Task Force (IETF) in 1996 under the title 'A Method for Web Robots Control'. According to this document a Web site administrator can indicate which parts of the site should not be visited by a robot, by providing a specially formatted file, in http://.../robots.txt. The other mechanism assumes that a Web author can indicate if a page may or may not be indexed, or analyzed for links, through the use of a special Hyper Text Markup Language (HTML) META tag i.e., a 'Robots META tag'. However, these mechanisms rely on cooperation from the robots, and are not even guaranteed to work for every robot. Moreover, as already suggested here above (DoS), some of these robots may not be so friendly. They could be run e.g., with the malicious intent of attacking a Web site (then, they just ignore the robots.txt file and the robots meta tags) so as it becomes overloaded and start refusing to serve legitimate users i.e., the human beings trying to use normally the site. Also, although the information made available on a site may not be confidential, an administrator may want to prevent an unlimited dissemination of it that would otherwise result of its indexing and referencing by all sorts of robots. The standard way of achieving this is to protect a Web site

through some form of authentication of which the more common method is to manage a list of registered users having a password so as they have to sign on upon accessing the site. The obvious drawback of this is that administrators must manage and update a closed list of users thus, requiring a registration step for a first consultation of a site also, assuming that users remember passwords in subsequent consultations. This may not be at all what administrator wanted to achieve in a first place and may even be counterproductive since it will certainly discouraged some individuals, willing to browse a site, to go further if they are requested to register.

## Object of the Invention

Thus, it is a broad object of the invention to prevent Web site contents from being investigated by robots.

It is a further object of the invention of not discouraging human beings, attempting to access a robot protected Web site, to proceed by imposing a registration at first access and a log on procedure at each subsequent access.

It is still another object of the invention not to rely on robots cooperation for barring them access to contents of Web sites.

Further objects, features and advantages of the present invention will become apparent to the ones skilled in the art upon examination of the following description in reference to the accompanying drawings. It is intended that any additional advantages be incorporated herein.

## Summary of the Invention

A method and a system for preventing robots from browsing a Web site beyond a welcome page are described. On receiving an initial request from an undefined originator Web site

responds to it with a welcome page including a challenge. Then, on receiving a further request from the undefined originator Web site can check whether the challenge is fulfilled or not. If fulfilled the undefined originator is assumed to be a human being and site keeps processing the further request and all subsequent ones if any. However, if challenge is not fulfilled the undefined originator is assumed to be a robot in which case all requests from that originator are not further processed.

The invention prevent Web site contents from being investigated by robots without requiring end users to register and site administrator to have to manage an access list of authorized users.

## Brief Description of the Drawings

**Figure 1** is an exemplary welcome page per the invention.

**Figure 2** shows the corresponding HTML code.

**Figure 3** shows the steps of the method of the invention.

**Figure 4** shows the further steps of the method when access to a Web site per the invention is denied, while a timer is on, for requests carrying a logged IP address.

**Figure 5** are other exemplary welcome pages with challenges.

## Detailed Description of the Preferred Embodiment

**Figure 1** illustrates the method according to the invention to prohibit robots from accessing a Web site beyond its welcome page. An exemplary welcome page as seen by an individual accessing a Web site e.g., at following URL [100] (Uniform

Resource Locator) 'http://www.anysite.com/welcome.html' is shown. Accessing to a Web site can be done with any available Web browser e.g., Netscape browser [110] from Netscape Communications Corporation, 501 E. Middlefield Road, Mountain View, CA 94043, the USA can be used. Then, according to this first method to prohibit robots, welcome page implements a dummy challenge that can simply be taken up by a human being while a robot should certainly fail it. Among various possibilities Figure 1 illustrates a typical challenge according to the invention. Welcome page thus shows an image [120] including, in this particular example, a few geometric forms that can be, unambiguously, referred to by a single word or expression in a language that the individual accessing the Web is assumed to be capable of reading. Then, associated to the image here including a square, a circle, a cross and a triangle, whoever is looking is prompted [130] to click e.g., on the cross [121]. A human being, desiring to go on and visit the site, will do it naturally while a robot will do nothing, or will do it wrongly simply because it just does not know what is a cross. Hence, this easily allows to discriminate a human being from a robot on the basis of their respective level of abstraction which is naturally high or very high for a human being while a robot is totally lacking this capacity. This allows to achieve the objectives of the invention which wants to prevent robots from browsing the site beyond the welcome page while neither imposing to the people accessing it the burden of having to register and to log on, nor requiring from the administrators of the site to have to manage a list of legitimate users.

To make the site even more resistant to browsing by a robot, that could be tailored to adapt to a given challenge or set of predictable such challenges, prompting can be made random so every time somebody comes in, the challenge is somehow different. For example, the cross can be moved to a different position on the image map so that the coordinates returned, when clicking on it, are different. Or, the

prompting can change in requesting e.g., to click on triangle instead.

**Figure 2** shows the source HTML (Hyper Text Markup Language) code [140] for this example which uses a server side map i.e., ismap [142], included in an anchor tag created with the <A..> ..</A> construct form [149] of the HTML language. Thus, when the user clicks on the cross [121], browser sends a request back to the server URL (/cgi-bin/challenge.exe) [141] including the X and Y coordinates of the click contained in ismap [142] so that the server can check that the click coordinates indeed matches the cross position. Anchor tag also carries an identification field i.e., id=XD34F739 [143] which is useful to correlate the answer, including the click coordinates, with the current challenge when this latter changes from one user to another as explained here above.

**Figure 3** depicts the steps of the method according to the invention when originator of an initial request to a Web site is responded with a challenge. Upon receiving this initial request [300] Web site server responds [310]. This is done through the establishment of a TCP connection with the originator (the reliable transport protocol of the Internet TCP/IP suite of protocols used by the Web). Response is in the form of a Web page including a challenge e.g., of the kind discussed in Figure 1. Then, having got server response, originator proceeds with a new transaction towards the Web site [320]. On receiving the new transaction Web site server checks if challenge is fulfilled [330]. If it is indeed the case [340] then, it assumes that originator is a human being and let it go. However, if Web server finds that challenge is not properly answered then, it must assume originator is a robot [350]. As a consequence, it stops processing current and further requests if any [351], which includes dropping the TCP connection or redirecting it to another site [352]. Also, the IP source address [361] may be remembered and a timer started

[362] so that the access to the site may be temporarily barred, from that IP source address, as explained in Figure 4.

**Figure 4** shows the case where the IP address of the originator is remembered when a robot is assumed. Then, one may decide, for a while, to drop or redirect immediately all requests issued with this particular source IP address, and all assumed to come from a robot (although this might not always be true since a robot may be behind a proxy or firewall performing a network address translation of all the IP source addresses it has to forward). Because IP source address of the request was logged and a timer started as explained in Figure 3, each time a new request is received [410] one first checks if the same source IP address is found [420]. If not, one may proceed normally [450]. If yes, timer is checked [430]. If it has elapsed, the logged IP address is reset [440] and new request is normally handled [450]. However, if timer has not elapsed, TCP connection is dropped or redirected [460] before resuming to a new received request [410].

**Figure 5** are other challenge examples that are easily answered by human being.

Figure 5-a takes the form of a quiz [510] which could be made as simple as shown [500] or as sophisticated as necessary to defeat elaborated robots or, alternatively, to adapt to a particular end-user population sharing a same type of skill.

Figure 5-b is another alternative combining images [520] and text [530] in an even more abstract way where the answer is suggested so is even better adapted to discriminate a human being from a robot. However, it is worth mentioning here that such a challenge is culture dependent and could serve as well to discriminate human beings on the basis of their social or ethnic origins.

## Claims:

What is claimed is:

1. A method for preventing robots from browsing a Web site beyond a welcome page [110], said method in said Web site comprising the steps of:

on receiving an initial request from an undefined originator:

responding to said initial request [300] with a said welcome page including a challenge [310];

on receiving a further request [320] from said undefined originator:

checking [330] whether said challenge is fulfilled or not;

if fulfilled:

assuming that said undefined originator is a human being [340];

keep processing said further request and subsequent ones if any [341];

if not:

assuming that said undefined originator is a robot [350];

stop processing said further request and subsequent ones if any [351].

2. The method according to claim 1 further including the steps of:

logging a source IP address [361] of said undefined originator

starting a timer [362].

3. The method according to any one of the previous claims wherein said step of stop processing said further request includes the further step of:

dropping or redirecting a TCP connection [352] established with said undefined originator.

4. The method according to any one of the previous claims further including, whenever receiving a new request [410], the steps of:

checking [420] whether a source IP address of said new request is matching said logged source IP address or not;

if matching:

checking [430] whether said timer has expired or not:

if expired:

resetting [440] said logged IP address; and

proceeding [450] normally with said new request;

if not expired:

dropping or redirecting said TCP connection [460];

if not matching:

proceeding [450] normally with said new request;

keep executing all here above steps with every new received request [410].

5. The method according to any one of the previous claims wherein said challenge includes prompting said undefined originator to perform a specific action [130].

6. The method according to any one of the previous claims wherein said prompting is different at each subsequent access of said Web site.

7. The method according to any one of the previous claims wherein said action includes having to make a choice among a plurality of options [120].

8. The method according to any one of the previous claims wherein said action calls for responding to a quiz [510].

9. The method according to any one of the previous claims wherein said action is suggested [530].

10. A system, in particular a Web site implementing a challenge access, comprising means adapted for carrying out the method according to any one of the previous claims.

11. A computer-like readable medium comprising instructions for carrying out the method according to any one of the claims 1 to 9.

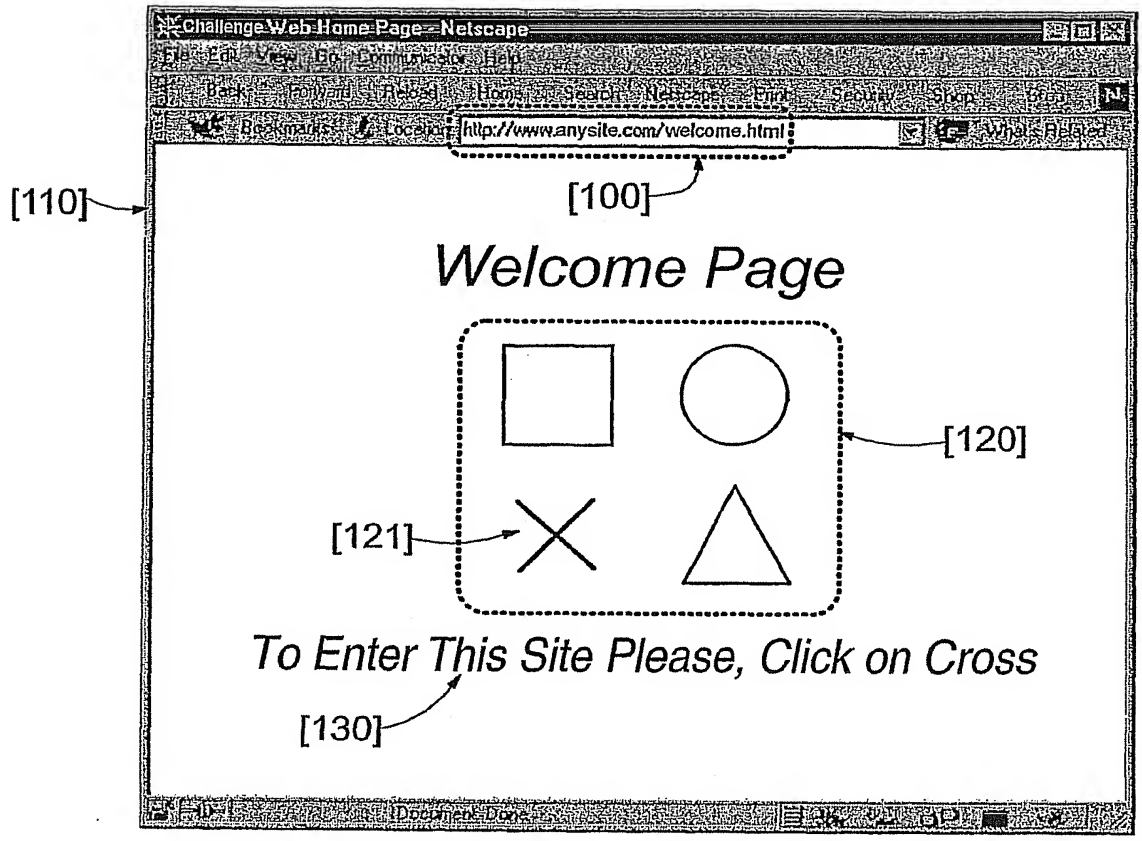



Figure 1

[140]

```
<!doctype html public "-//w3c//dtd html 4.0 transitional//en">
<html><head><title>Cross</title></head>
<body text="#000000" bgcolor="#000000" link="#0000EE" vlink="#551A8B" alink="#FF0000">

<h1><i><font color="#3366FF">Welcome Page</font></i></h1></center>

<hr WIDTH="100%"><br>                                        [142]
<a href="/cgi-bin/challenge.exe?id=XD34F739"><img SRC="cross.gif" ismap ...
                                                    height=135 width=183></a>
</body>
</html>
```

[141] [143] [149]

Figure 2

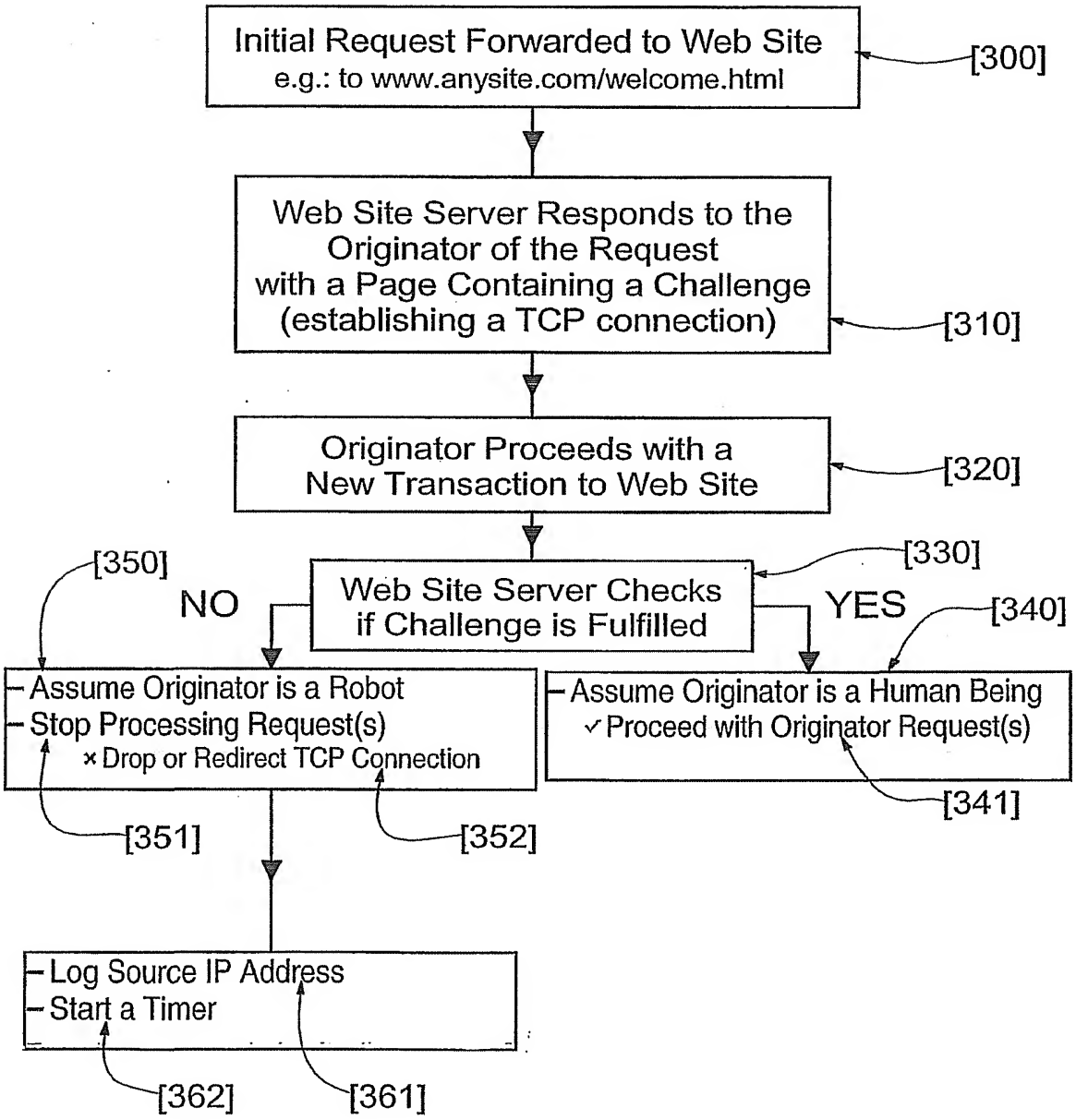

Initial Request Forwarded to Web Site
e.g.: to www.anysite.com/welcome.html
[300]
Web Site Server Responds to the Originator of the Request with a Page Containing a Challenge (establishing a TCP connection)
[310]
Originator Proceeds with a New Transaction to Web Site
[320]
[330]
[350]
NO
Web Site Server Checks if Challenge is Fulfilled
YES
[340]
Assume Originator is a Robot
Stop Processing Request(s)
× Drop or Redirect TCP Connection
Assume Originator is a Human Being
✓ Proceed with Originator Request(s)
[341]
[351]
[352]
Log Source IP Address
Start a Timer
[362]
[361]

Figure 3

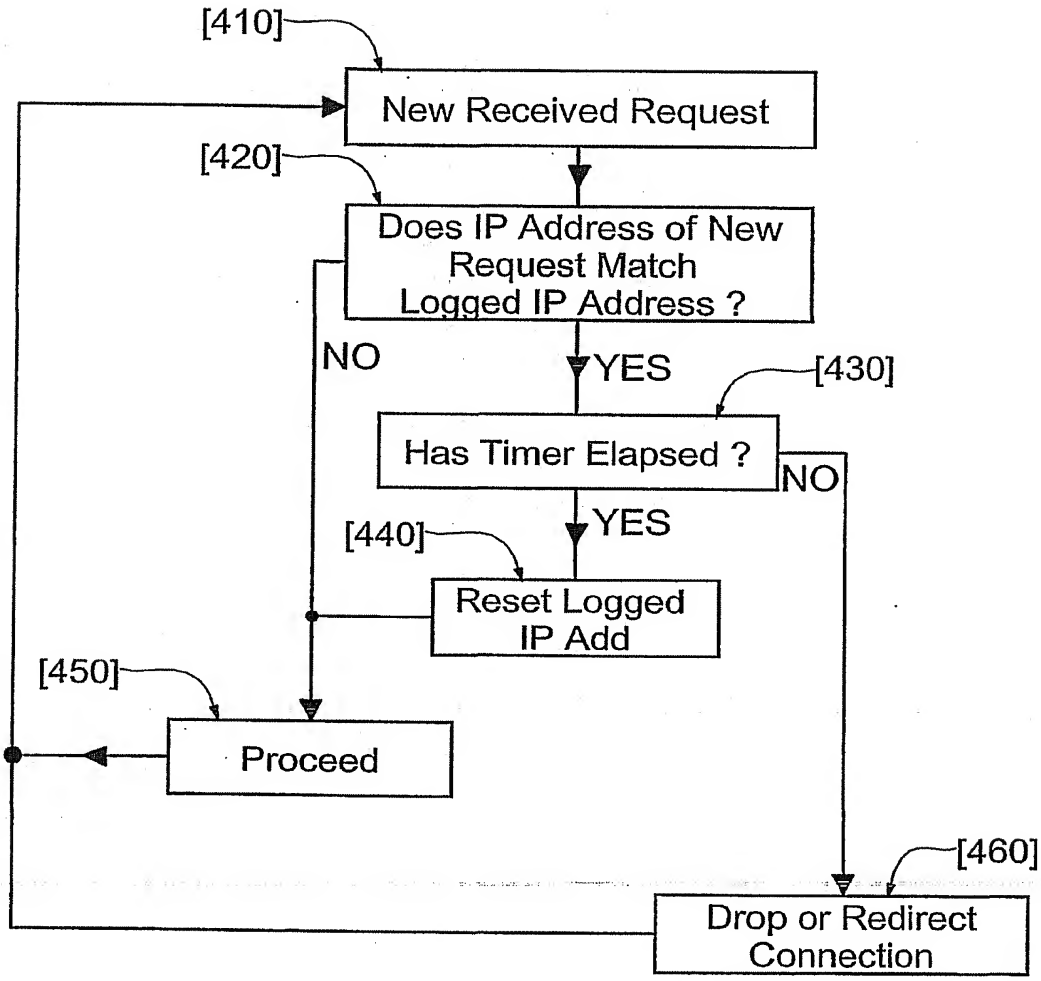

[410]
New Received Request
[420]
Does IP Address of New Request Match Logged IP Address ?
NO
YES
[430]
Has Timer Elapsed ?
NO
YES
[440]
Reset Logged IP Add
[450]
Proceed
[460]
Drop or Redirect Connection

Figure 4


Figure 5-a

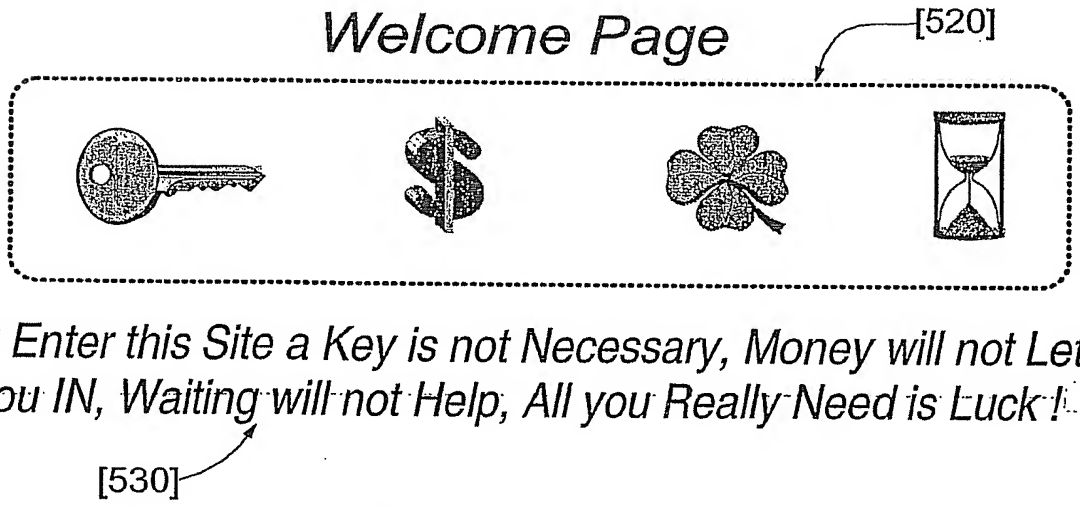



Figure 5-b